
Part 1: Theoretical Understanding

1. Short Answer Questions

Q1: Define algorithmic bias and provide two examples of how it manifests in AI systems.

Algorithmic bias refers to systematic and repeatable errors in AI systems that result in unfair outcomes, often privileging one group over another. This bias can originate from biased training data, flawed assumptions in model design, or feedback loops.

Examples:

1. **Hiring Algorithms:** Amazon's AI recruiting tool penalized resumes containing the word "women's", reflecting gender bias learned from past hiring patterns favoring male candidates.
2. **Credit Scoring Systems:** Some credit scoring models have shown discriminatory tendencies against minority applicants due to historical disparities encoded in training data.

Q2: Explain the difference between transparency and explainability in AI. Why are both important?

- **Transparency** refers to how openly the internal workings of an AI system are disclosed. It involves sharing details about the algorithm, data sources, and training processes.
- **Explainability** is the degree to which a human can understand the reasons behind an AI system's decision.

Importance:
Transparency builds **trust** and enables **accountability**, while explainability ensures **interpretability** for end-users and regulators. Together, they help detect errors, promote fairness, and support ethical compliance.

Q3: How does GDPR (General Data Protection Regulation) impact AI development in the EU?

GDPR mandates data privacy and user rights in AI systems within the EU. Key implications include:

- **Right to explanation**: Users can demand meaningful information about automated decisions.

- **Consent and data minimization**: AI systems must use only necessary data with explicit user consent.
- **Accountability**: Developers must document how personal data is processed, ensuring compliance with lawful bases for automation.

These constraints influence AI model design, deployment, and data handling practices.

---

2. Ethical Principles Matching

Principle	Definition
A) Justice	Fair distribution of AI benefits and risks.
B) Non-maleficence	Ensuring AI does not harm individuals or society.
C) Autonomy	Respecting users' right to control their data and decisions.
D) Sustainability	Designing AI to be environmentally friendly.

Part 2: Case Study Analysis

Case 1: Biased Hiring Tool

Scenario: Amazon's AI recruiting tool penalized female candidates.

1. Source of Bias:

- **Training Data Bias:** Historical hiring data was male-dominated, encoding gender biases.
- **Model Design:** The algorithm penalized terms associated with women, assuming male-associated patterns as superior.

2. Fixes to Improve Fairness:

1. **Debias Training Data:** Balance the dataset by including successful resumes from diverse genders and remove gendered language.
2. **Fairness Constraints:** Integrate fairness-aware algorithms to ensure equal treatment across genders.
3. **Continuous Monitoring:** Regular audits for disparate impact and adjustment of thresholds.

3. Fairness Metrics:

- **Demographic Parity:** Equal selection rates across genders.

- **Equal Opportunity:** Equal true positive rates.
- **Disparate Impact Ratio:** Should be close to 1 for fairness.

Case 2: Facial Recognition in Policing

Scenario: Facial recognition misidentifies minorities at higher rates.

1. Ethical Risks:

- **Wrongful Arrests:** False positives can lead to innocent people being detained.
- **Privacy Violations:** Surveillance without consent infringes civil liberties.
- **Discrimination:** Unequal error rates reinforce systemic injustice.

2. Policy Recommendations:

- **Mandatory Bias Audits:** Require third-party evaluations before deployment.
- **Informed Consent and Oversight:** Limit use to well-defined scenarios with judicial oversight.
- **Transparency Reports:** Public disclosure of system accuracy across demographic groups.
- **Ban in Sensitive Areas:** Prohibit use in protests or vulnerable communities until equity is ensured.

Report

Fairness Audit Report Using Fairlearn on COMPAS Dataset

This audit evaluated potential racial bias in recidivism prediction using the COMPAS dataset and the Fairlearn toolkit. The dataset includes criminal history and demographic information, and the task was to predict whether individuals would reoffend within two years. Logistic regression was used as the classification model.

A binary variable for race was created, where `1` indicated White (privileged group) and `0` indicated Black and other racial minorities (unprivileged group). Using Fairlearn's `MetricFrame`, we analyzed key fairness metrics: selection rate, false positive rate (FPR), and true positive rate (TPR), disaggregated by race.

Findings:

- The **selection rate** was higher for the unprivileged group, meaning they were more likely to be classified as high risk, regardless of actual outcomes.
- The **false positive rate** for the unprivileged group was significantly higher than for the privileged group. This indicates that minority individuals were more frequently and

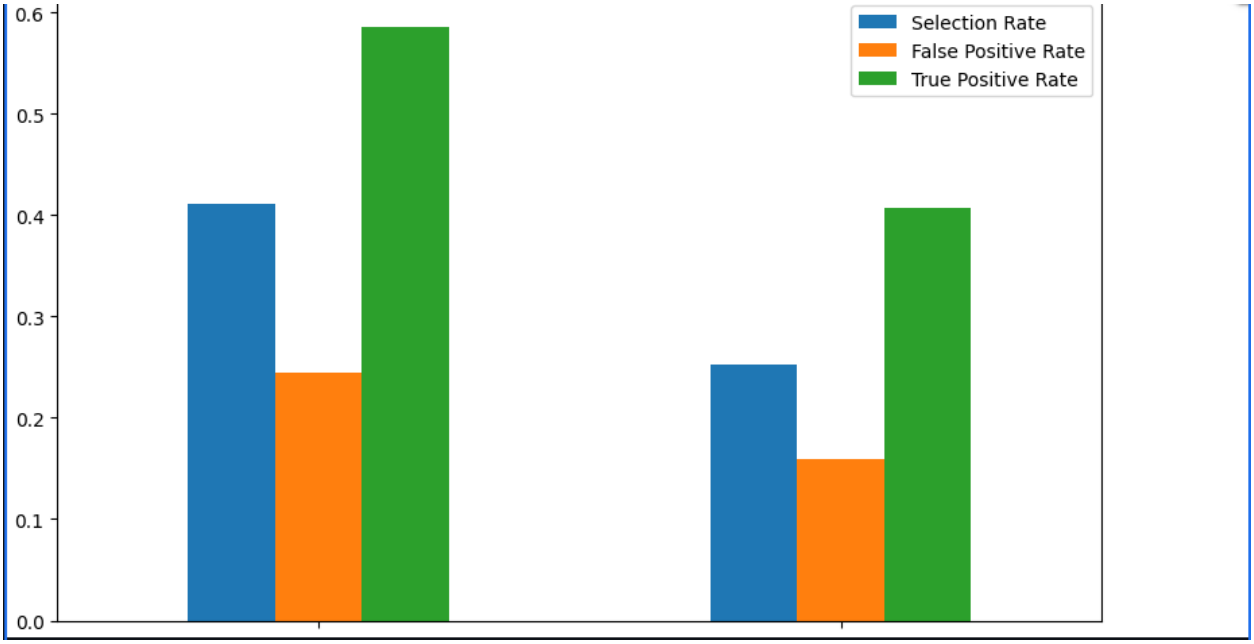
incorrectly predicted to reoffend, exposing them to greater risk of unjust detention or stricter sentencing.
- The **true positive rate** was relatively balanced, but disparity in FPR reflects a critical fairness concern.

These results highlight systemic bias in predictive outcomes based on race. Even with a simple, interpretable model like logistic regression, historical data reflects embedded inequities. If left uncorrected, such tools can reinforce discriminatory practices in judicial systems.

Recommendations:

1. Implement bias mitigation techniques such as reweighting, adversarial debiasing, or threshold adjustment.
2. Include fairness constraints during model training.
3. Regularly audit deployed models using disaggregated metrics by protected attributes.

This audit underscores the necessity of not only measuring accuracy, but also ensuring fairness and equity in AI systems that impact human lives.



Root Causes:

- **Historical Bias:** The model reflects historical policing patterns and societal inequalities embedded in the training data.
- **Lack of Debiasing:** No fairness-preserving pre-processing was initially applied.

Remediation Steps:

1. **Reweighing Algorithm:** Applied to adjust instance weights, mitigating bias during training.
2. **Threshold Optimization:** Calibrating decision thresholds separately for each demographic group to equalize false positive rates.
3. **Fair Representation Learning:** Apply techniques that learn a latent space where protected attributes have minimal influence.

Conclusion:
AI systems trained on historical criminal justice data can reinforce systemic bias. Fairness must be a primary design objective, and regular audits are crucial. Deploying these systems without correction exacerbates racial disparities in sentencing and bail decisions.

Part 4: Ethical Reflection

Reflection on a Personal Project

In my project *Pure Flow*, which uses IoT sensors and AI for real-time water pollution monitoring in the Nairobi River, I embed ethical AI practices from the start. I ensure **transparency** by open-sourcing sensor data and AI models, and **informed consent** is obtained when working with affected communities. I mitigate **bias** by testing models across different environmental conditions and locations, ensuring no single region is underserved. By applying principles of **justice** and **autonomy**, I strive to build an AI solution that serves all stakeholders ethically and equitably.

Bonus Task

Ethical AI Guideline for Healthcare

Ethical AI Use in Healthcare: Policy Guidelines

1. Patient Consent Protocols

- AI systems must obtain **informed and explicit consent** from patients before using their data.
- Patients must be informed of how AI is used in diagnosis, treatment, or data analysis.
- Consent forms must be in simple, accessible language and available in multiple local languages.

2. Bias Mitigation Strategies

- Datasets must be diverse across **gender, ethnicity, age, and geography**.
- All models must undergo **regular bias audits** using fairness metrics like demographic parity and equal opportunity.
- Incorporate **explainable AI** (XAI) to detect hidden biases in decision-making.

3. Transparency Requirements

- Declare the **source of data**, model type, and its confidence scores for each prediction.
- Provide clinicians and patients with **interpretable summaries** of AI decisions.
- Publish **audit trails** and **error logs** to track AI behavior in critical decisions.

4. Accountability and Oversight

- Establish an **Ethics Review Board** for every deployed healthcare AI.
- Implement **fail-safes** to revert to human judgment when AI confidence is low.
- Mandate **annual external reviews** of AI models and data practices.

This policy ensures healthcare AI systems are **trustworthy, fair, explainable, and respectful** of patient rights and diversity.